Computational Statistics and Data Analysis 56 (2012) 664–679

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Sample distribution function based goodness-of-fit test for complex surveys

Jianqiang C. Wang

Hewlett-Packard Labs, Palo Alto, CA, 94304, United States

ARTICLE INFO

Article history:
Received 11 April 2010
Received in revised form 5 September 2011
Accepted 6 September 2011
Available online 21 September 2011

Keywords:
Anderson–Darling test
Convergence in functional space
Kolmogorov–Smirnov test
Gaussian process
Rao–Kovar–Mantel estimator

ABSTRACT

Testing the parametric distribution of a random variable is a fundamental problem in exploratory and inferential statistics. Classical empirical distribution function based goodness-of-fit tests typically require the data to be an independent and identically distributed realization of a certain probability model, and thus would fail when complex sampling designs introduce dependency and selection bias to the realized sample. In this paper, we propose goodness-of-fit procedures for a survey variable. To this end, we introduce several divergence measures between the design weighted estimator of distribution function and the hypothesized distribution, and propose goodness-of-fit tests based on these divergence measures. The test procedures are substantiated by theoretical results on the convergence of the estimated distribution function to the superpopulation distribution function on a metric space. We also provide computational details on how to calculate test p-values, and demonstrate the performance of the proposed test through simulation experiments. Finally, we illustrate the utility of the proposed test through the analysis of US 2004 presidential election data.

1. Introduction

Testing the parametric distribution of a random variable or vector is a fundamental problem in both exploratory data analysis and inferential statistics. Classical test procedures include the Pearson Chi-squared test for discrete data, Kolmogorov–Smirnov test, Anderson–Darling test and Cramér–von Mises test for continuous variables, among others. The Kolmogorov–Smirnov test, Anderson–Darling test, and Cramér–von Mises test are based on the discrepancy between the empirical distribution function (EDF) and the hypothesized distribution, and are often referred to as EDF-based tests.

The theoretical foundation of EDF-based goodness-of-fit tests is the convergence of EDF, viewed as a random element on a metric space, to the true distribution over the support of the study variable. Functional central limit theorems have been established for EDF and its variations (e.g., Donsker, 1952). A key assumption for the classical weak convergence result is that the observations are independent and identically distributed (*i.i.d.*). The *i.i.d.* assumption is usually violated when data are obtained from a large-scale complex survey. Thus, novel statistical theory is required for migrating the EDF-based inference to the sampling context. In this paper, we introduce the classical EDF theory to survey sampling, derive theoretical results for design weighted distribution functions, and demonstrate how these results can be exploited to construct goodness-of-fit tests.

The effects of sampling design (for example, stratification and clustering) on Pearson Chi-squared test and Wald's test for categorical data have been well studied in the literature (Cohen, 1976, Fellegi, 1980, Holt et al., 1980, Rao and Scott, 1981, among others). To perform tests on a continuous variable, one can in principle discretize the variable and then apply the Pearson chi-squared test to the discretized variable. The discretization procedure is subject to the choice of user-defined cutoff values, and the test result is affected by this subjective choice. Krieger and Pfeffermann (1997) proposed

E-mail addresses: qqwjq9916@gmail.com, jianqiang.jay.wang@hp.com.

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.09.015

an alternative goodness-of-fit test for continuous survey variables. The authors viewed the unweighted sample probability density function (PDF) as a "distorted" version of the superpopulation density, and derived the sample PDF based on a fully specified superpopulation distribution, coupled with the estimated relationship between inclusion probability and the study variable. The Krieger–Pfeffermann test suffers the following problems:

1. The test is designed for testing against a fully-specified null distribution, and extending the procedure to testing a parametric family is not straightforward.
2. The "extracted" distribution of sample data usually does not have a closed-form expression, and has to be derived on a case-by-case basis.
3. The test requires to approximate the conditional expectation (or log-conditional expectation) of inclusion probabilities by a polynomial function of study variable, and the approximation errors are ignored in further analysis.

The focus of our research is to propose a general procedure for testing the superpopulation distribution of a survey variable. An attractive test should be readily applicable to a large number of parametric families, flexible enough to incorporate auxiliary variables if available, and accommodate replication variance estimation. Our proposed test enjoys all the stated advantages and is justified theoretically in Section 3.

The proposed method begins with the sample distribution function (SDF), an estimator of the population distribution. The estimation of distribution functions in complex surveys has been extensively explored in literature. Dunstan and Chambers (1986) offered a model-based perspective and proposed an estimator under the ratio model. Rao et al. (1990) furnished a thorough treatment of estimators of distribution functions. In Rao et al. (1990), the authors proposed a ratio estimator, a difference estimator and an estimator that is asymptotically both design-unbiased and model-unbiased (often referred to as the RKM estimator). Asymptotic normality of distribution function and quantile estimators have been shown by Francisco and Fuller (1991) under stratified cluster sampling. The properties of the Dunstan and Chambers estimator and the RKM estimator were further investigated by Chambers et al. (1992). Dorfman (2009) reviewed the survey literature on estimating distributions and quantiles. In the present paper, we start with design-based distribution function estimators, including the Horvitz–Thompson (HT) estimator, Hájek estimator and RKM estimator, and construct goodness-of-fit tests based on these estimators.

The remainder of the paper is organized as follows. The proposed test is elaborated in Section 2, followed by theoretical arguments on SDF in Section 3. Next, we generalize the test to two-sample goodness-of-fit testing situations in Section 4, and provide computational details for test implementation in Section 5. Then, we demonstrate the performance of proposed test through simulations in Section 6, and illustrate the effectiveness of our test procedure through the analysis of a real dataset, specifically the US 2004 presidential election data. Finally, Section 8 contains some concluding remarks.

2. SDF-based goodness-of-fit test

In this section, we describe the proposed procedure for testing whether the distribution of a univariate variable Y belongs to a parametric family or not, based on a sample of observations from the finite population. Extending the proposed methodology to multivariate setting is straightforward and thus omitted from the current paper.

We work under the superpopulation setup of Rubin-Bleuer and Kratina (2005), in which the authors elaborated the "superpopulation viewpoint" proposed by Hartley and Sielken (1975). Consider a finite population $\mathcal{U}_N$, in which each population element is associated with a variable of interest y, a q-dimensional vector of design variables $\boldsymbol{d}$, and a k-dimensional vector of auxiliary variables $\boldsymbol{x}$. For notional convenience, it is assumed that the finite population $\mathcal{U}_N$ is of size N and the collection of population indices is $U_N = \{1, \ldots, N\}$. The superpopulation associated with $\mathcal{U}_N$ is embedded with a probability space $(\Omega, \mathcal{F}, \xi)$, and the collection of random variables $(y^N, \boldsymbol{d}^N, \boldsymbol{x}^N)$ are ξ-measurable. More rigorously, one should write $y_i = y_i(\omega)$, $\boldsymbol{d}_i = \boldsymbol{d}_i(\omega)$, $\boldsymbol{x}_i = \boldsymbol{x}_i(\omega)$, with $\omega \in \Omega$, but we will suppress ω in $(y_i(\omega), \boldsymbol{d}_i(\omega), \boldsymbol{x}_i(\omega))$ throughout the paper for compactness.

A random sample $A \subset U_N$ is the collection of indices of selected elements. Let $\mathcal{A}$ denote the set of all possible samples under the sampling design, hence sample A is a random element in $\mathcal{A}$. The size of sample A is often a random variable, and denoted as n. According to Rubin-Bleuer and Kratina (2005) and the references therein, conditioning on the design variables $\{\boldsymbol{d}_i\}_{i=1}^N$, the sampling design p is a mapping from $\mathcal{A}$ to the unit interval [0, 1]. Coupled with the superpopulation distribution for design variable, we define the product space $(\mathcal{A} \times \Omega, \sigma(\mathcal{A}) \times \mathcal{F})$ where $\sigma(\mathcal{A})$ is the power set of $\mathcal{A}$, and regard the sampling design as a random quantity defined on the product space. Any sample-based estimator $\widehat{\boldsymbol{\theta}}$ is viewed as a random variable on the product space that depends on both $A \in \mathcal{A}$ and $\omega \in \Omega$. The estimator $\widehat{\boldsymbol{\theta}}$ depends on ω through the population variables $(y^N(\omega), \boldsymbol{d}^N(\omega), \boldsymbol{x}^N(\omega))$.

The superpopulation distribution is initially hypothesized to come from a family of parametric distributions $\{F(y; \boldsymbol{\theta}) | \boldsymbol{\theta} \in \Theta\}$, and the goal is to test the following hypothesis

$$H_0 : F(y) \in \{F(y; \boldsymbol{\theta}) | \boldsymbol{\theta} \in \Theta\} \quad \text{vs.} \quad H_a : F(y) \notin \{F(y; \boldsymbol{\theta}) | \boldsymbol{\theta} \in \Theta\}, \tag{1}$$

based on a random sample from the finite population. The design generating the sample could be simple random sampling with or without replacement, Poisson sampling, stratified sampling or other designs that satisfy certain regularity conditions elaborated later. The first-order inclusion probability for the i-th population element is denoted as $\pi_i = \pi(\boldsymbol{d}_i)$, for some